Development of a Faculty Research Interest Resource

Jerry Fowler*, David A. Wheeler, Pat W. Camerino,
Olga Bat, Paula E. Burch
Baylor College of Medicine
Houston, Texas

We have developed a faculty research interests resource by "mining" MEDLINE for relationships that are not directly queryable through the normal MEDLINE schema. Faculty citations are retrieved and World-Wide Web pages built to interconnect authors, their citations, and the MeSH terms that have been assigned to these citations. The design and development of the resource are discussed and examples of the results illustrated.

INTRODUCTION

The customs, habits and work routines of large medical institutions make it difficult or impossibile to keep track of centers of expertise outside of one's field. In spite of internal newsletters and other promotional material that bring visibility to selected projects, obtaining timely information on "who does what" can be daunting. Assistance to faculty networking can be extremely important for a variety of reasons. Often, ready sources of academic expertise exist close at hand and can be tapped for advice relating to grants, publications, and technical information. Information about who is doing what can promote collaboration and other collegial exchanges that are beneficial to an individual researcher or clinician as well as serving the higher purposes of the institution itself.

Since 1981, Baylor College of Medicine (BCM) has maintained a faculty database that tracks faculty achievement in areas of publication and research awards. For the last five years, the database has tracked faculty research interests as well. The database, in relational format, was developed in Paradox by one of us (P.C.). This resource is maintained at considerable effort through annual survey of faculty productivity (83 percent compliance requiring a two month collection effort) and manual updating of the data. The information, used primarily for reports to the Executive Officers, is potentially available to all faculty simply by calling the Office of the President.

Several benefits have been realized by the use of this database in the administrative and executive offices of the College:

- It was used retrospectively to assess the effectiveness of changed strategic emphases mandated by the College in 1987.
- It can assist the Office of Public Affairs with journalist media queries to identify faculty with particular expertise.
- It has assisted current strategic planning by BCM executives who have identified eight new areas of emphasis. The database was queried to establish Faculty committees assembled from local experts in these fields to make recommendations for ways to expand programs in these areas.

Use of this database can yield other benefits. For a variety of reasons, however, this information resource remains underutilized. It has had little chance to serve its intended purpose which is to promote exchange of information among the faculty at the College. There are several disadvantages to the current system:

- Faculty input is indirect, and faculty members' descriptions of their own work are often inadequate. Vague terms such as "research," "study," and "analysis" appear frequently, and individual variations in nomenclature can cause overlapping interests to be missed.
- Data collection and update is an arduous process, paper distribution is expensive, and the database is difficult to access otherwise.

A faculty committee on Information Technology at Baylor has recommended that a resource like

**Author's current address: MCC, Austin, TX*

0195—4210/96/$5.00 © 1996 AMIA, Inc.

the interests database be made available via the World–Wide Web (or, simply "the Web"). The committee has cited difficulty of access and incompleteness of data as inadequacies of the current system. For these reasons we have undertaken a development effort to augment the current system with one that addresses the aforementioned problems.

To provide an already well-established indexing scheme, we are using the Medical Subject Headings (MeSH) of MEDLINE and the MeSH hierarchy as derived from the Unified Medical Language System (UMLS) Metathesaurus [1], to construct a web of pages that relate BCM authors to the MeSH terms that were used to index their published work. Using this resource, one can determine which BCM researchers have written something on the topic of a given MeSH heading.

Related Work

The idea for this work developed during previous work on "data mining" in MEDLINE for the purpose of automatic categorization of Web pages [2, 3]. The work draws on the concept of semantic locality in the UMLS [4]. The faculty research interest resource assembles authors into semantic locales based on the MeSH headings assigned to their published work. Another significant exploitation of semantic locality using the UMLS is the work of Yang and Chute on automatic text categorization [5, 6].

In a project with motivations similar to ours, Schwartz grouped computer users with similar interests or expertise [7] by collecting samples of electronic mail from 15 sites around the Western Hemisphere and building a graph based on senders and receivers in order to compute an *interest distance* between individuals.

METHOD

Our approach to building a useful index of faculty research interests relies on the wealth of information that is available in MEDLINE. MEDLINE contains several logical data relations that are not directly accessible by query. The data are "mined" to extract these relations and make use of them. The method is straightforward: We retrieve from MEDLINE the citations for which the institution is recorded as "Baylor College of Medicine," and use the results to build a web of pages containing citations, authors, and MeSH terms. The pages produced are accessible via the Web, and provide a number of interesting results.

Object–based design was used to develop this resource; that is, while the objects developed in the program did not inherit properties from supertypes (object orientation), each object had a handling mechanism that dealt with creation and incremental modification (no deletion mechanism was provided, since this is a historical resource) in a manner analogous to object methods.

Inputs

In order to provide a model for development of a college research information resource, we incorporated data from several sources into this database.

UMLS Metathesaurus. Two relations from the UMLS Metathesaurus were used to build the web of MeSH terms. The concept relation (`mrcon`) associates the terms assigned to individual citations with the appropriate Metathesaurus unique identifier, and the context relation (`mrcxt`) provides the hierarchical structure within the MeSH.

MEDLINE. Citations were collected from a local copy of MEDLINE using the query `INSTITUTION = (Baylor College of Medicine)`. Fields retrieved included author, title, institution (the institution field frequently contains departmental information that can help differentiate authors), source, MeSH headings, and abstract. The unique identifier is also retrieved to serve as a pathname for the citation's URL (Universal Resource Locator).

College Resources. The database maintained by the College's Faculty Resources provides information about BCM authors including full name, degree, position, and office. The personnel database provides contact information including email address, telephone, and fax.

Processes

The procedures developed for this project were comparatively simple, as befits the simplicity of the data mining involved. Each citation was identified by its unique mesh identifier, each MeSH term by its Metathesaurus identifier, and each author by a string derived from the MEDLINE citation standard for authors. This enabled the construction methods for the various objects to build URL's for other objects knowing only a path and identifier, without necessarily requiring that the object exist at the time the URL was created.

Author Mapper. The author mapper constructs a mapping from MEDLINE's `surname--initials` style of naming to the full names by which the authors are known to the institution. This task is better performed in advance of the creation of the faculty interests resource, because automating the task is open to several obvious difficulties: Surnames can appear in several forms; in our sample from MEDLINE, there were three different spellings of the name "De Bakey," and at least one other name was present with and without the suffix, "Jr." The same individual can appear with one or two initials. Other problems occur, including typographical errors. Consequently, human inspection of the results of this mapping is essential.

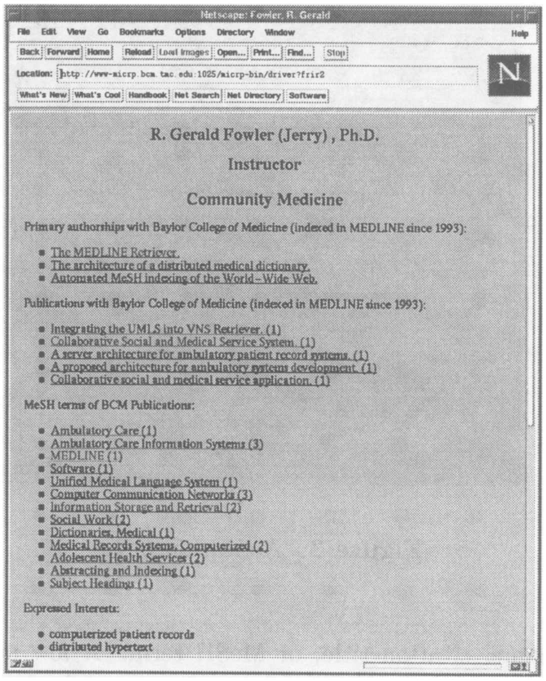
Term Webmaker. The `term webmaker` is invoked for each main MeSH heading in each citation. A new page is created for the term if it does not already exist, and pointers are created to each parent node found in the MeSH source of UMLS Metathesaurus (there may be parents in multiple contexts). The `term webmaker` is then recursively invoked for each of these parents. A child link to the page that invoked it is added to each page that is recursively invoked. The resulting acyclic directed graph contains exactly that portion of the MeSH hierarchy that contains all the MeSH terms used in citations authored by colleagues at BCM.

Author Webmaker. The `author webmaker` is invoked for each author named in each citation, creating a new page for the author if appropriate, and identifying the author by full name, department, and several other points of interest. Links are created to the corresponding citation page, to the MeSH terms, and to co-authors' pages. A link is created from each MeSH term to the author, and the ancestors of those MeSH terms are each updated to reflect the presence of an additional author in the tree below them.

Citation Webmaker. The `citation webmaker` is the simplest of all these constructors. It involves merely reformatting the information available from the MEDLINE citation into HTML, placing links to the author pages and to the MeSH pages. As a citation is created, a pointer to it is also added to the appropriate MeSH pages, and their ancestors are updated to reflect the presence of an additional citation.

Outputs

The outputs that result from this set of processes resemble Figures 2 through 3, found in the following pages. Faculty pages (as in Figure 1) link authors to their interests (as recorded in MEDLINE), co-authors, and to those papers published in MEDLINE with BCM as the lead institution (including papers authored in collaboration with other institutions is a non-trivial problem, probably best solved by query to the authors themselves). MeSH pages link the relevant parts of the MeSH hierarchy to the faculty members who have published on those topics and to the papers themselves. An example mesh page appears in Figure 2. Figure 3 illustrates an example citation page. Citation pages make explicit each relation that was drawn elsewhere between the MeSH hierarchy and the faculty.



Location: http://www-micrp.bcm.tmc.edu:1025/micrp-bin/driver?frir2

R. Gerald Fowler (Jerry) , Ph.D.

Instructor

Community Medicine

Primary authorships with Baylor College of Medicine (indexed in MEDLINE since 1993):

- The MEDLINE Retriever.
- The architecture of a distributed medical dictionary.
- Automated MeSH indexing of the World–Wide Web.

Publications with Baylor College of Medicine (indexed in MEDLINE since 1993):

- Integrating the UMLS into VNS Retriever. (1)
- Collaborative Social and Medical Service System. (1)
- A server architecture for ambulatory patient record systems. (1)
- A proposed architecture for ambulatory systems development. (1)
- Collaborative social and medical service application. (1)

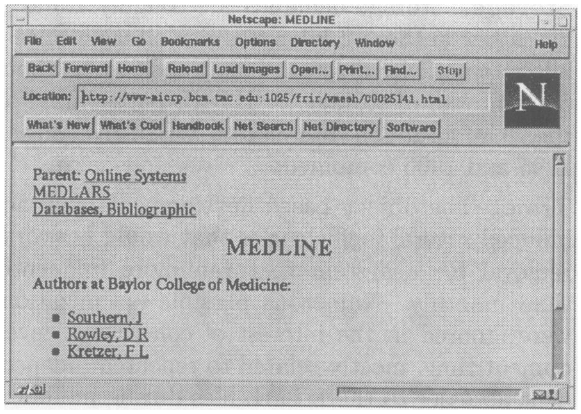
MeSH terms of BCM Publications:

- Ambulatory Care (1)
- Ambulatory Care Information Systems (3)
- MEDLINE (1)
- Software (1)
- Unified Medical Language System (1)
- Computer Communication Networks (3)
- Information Storage and Retrieval (2)
- Social Work (2)
- Dictionaries, Medical (1)
- Medical Records Systems, Computerized (2)
- Adolescent Health Services (2)
- Abstracting and Indexing (1)
- Subject Headings (1)

Expressed Interests:

- computerized patient records
- distributed hypertext

Figure 1 An author page



Location: http://www-micrp.bcm.tmc.edu:1025/frir/wmesh/C0025141.html

Parent: Online Systems
MEDLARS
Databases, Bibliographic

MEDLINE

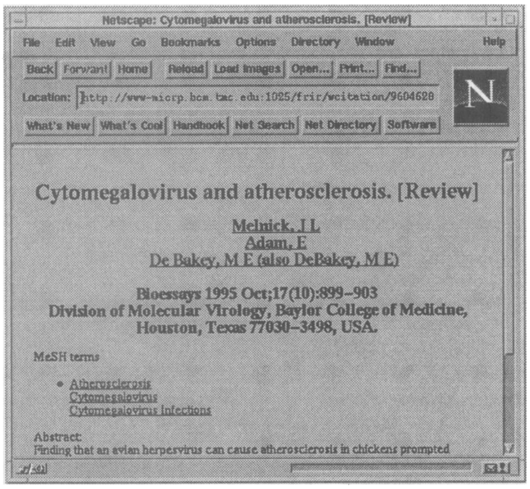
Authors at Baylor College of Medicine:

- Southern, J
- Rowley, D R
- Kretzer, F L

Figure 2 A MeSH page

RESULTS

The Faculty Research Interest Resource is now accessible to colleagues at BCM. Entry into the resource is a query page allowing search by author



Location: http://www-micrp.bcm.tmc.edu:1025/frir/wcitation/9604628

Cytomegalovirus and atherosclerosis. [Review]

Melnick, J L
Adam, E
De Bakey, M E (also DeBakey, M E)

Bioessays 1995 Oct;17(10):899–903
Division of Molecular Virology, Baylor College of Medicine,
Houston, Texas 77030–3498, USA.

MeSH terms

- Atherosclerosis
  Cytomegalovirus
  Cytomegalovirus Infections

Abstract:
Finding that an avian herpesvirus can cause atherosclerosis in chickens prompted

Figure 3 A citation page

name, citation title, or MeSH term. A few particulars about the system performance in terms of storage, execution time, and human effort are worthy of note.

Storage. Storage demands are roughly 1.5 kilobytes per author, 2 kilobytes per citation, and 1 kilobyte per MeSH term. This translates to about 6 megabytes of storage for HTML pages based on the BCM citations entered into MEDLINE during 1995 and 1996 combined.

Time. The object–based development approach induced several inefficiencies that would be worth removal for a system that ran more frequently than monthly. Numerous possible optimizations were ignored in the interest of conserving development time, mostly related to repeated independent accesses to the HTML files during modification. The system was developed in `Tcl`, which is a programmer–friendly language, but tends to run slowly. Consequently, it appears to take roughly 30 seconds of clock time to add a citation to the web. This is tolerable only in the context of a process that is run seldom, on relatively small amounts of data (roughly 100 citations per month), as this one is.

Human effort. Little time or human effort will be required for monthly update. Our design makes this process straightforward. On the order of 100 BCM citations are added to MEDLINE each month; we will be able to employ the retrieval component of the MeSH-Indexer medibot [2] with the query `INSTITUTION = (Baylor College of Medicine) AND (ENTRYMONTH =` *last month*`)` as the first step of an automated process to accrue the newest MEDLINE citations and add them to the research interest web. The computation will require about 50 minutes of real time by our estimates. Examination of the resulting logs for inconsistent author names and the like will require additional time on the part of a human analyst.

DISCUSSION

There are several factors that present difficulty in the development of the faculty interest resource. First, only those publications that are indexed in MEDLINE can contribute to the MeSH hierarchy. Second, only those citations for which a BCM collaborator is first author are retrieved using our simple query. A more complex query sequence that systematically retrieved all citations for each author would likely run afoul of the limited identification scheme adopted in MEDLINE by which only first and middle initials are stored. Another issue is the question of reconciling the maturation of vocabulary over the years. We have not yet evaluated the impact of using the 1996 UMLS for our concept graph.

An enhancement to the resource would be to add other information from the Metathesaurus to the MeSH pages, such as the definition. At this time, we consider it inappropriate to duplicate this information, which is available from the UMLS server at the National Library of Medicine (NLM). We may incorporate links to the NLM server if demand is felt.

The use of HTML as the storage medium, as opposed to using conventional or object–oriented database technology, was motivated for several reasons. First, the output can be manually inspected without difficulty and displayed and tested trivially using a Web browser. Second, serving static HTML is less of a load on the server than the computation that would be necessary to generate these pages on the fly from a database gateway interface. In addition, because this is a largely historical resource, it changes slowly and only through accretion. Consequently, the greatest benefits of using database technology, multiple access control and management of deletion and modification, are wasted. Finally, some of the information stored would be prohibitive to compute on the fly, so the benefit of formal database structure is further diminished.

Disadvantages of using HTML include high storage overhead, which we chose to sacrifice for speed of delivery. More important, the flexibility of the application is reduced, because automatic query becomes a matter of parsing HTML to reveal the underlying schema. This is an inherent limitation of the Web, which serves well as a user interface, but less so as a resource for computation.

A useful enhancement would be to generalize the description of an author's interests by identifying terms that are siblings or cousins and assigning their least common ancestor to the author's page.

Anticipated Benefits

Future uses of a Web–based resource in the College will be several, and will stem in part from the wide accessibility of the resource.

- Grant programs could be proactively targeted to researchers involved in particular fields by automatically combing the grants database and the faculty interests resource to match grant programs to specific faculty members and notify them of opportunities.
- Abstracts could more easily be routed to the people most likely to be able to review them.
- Identifying the interests of the faculty could be useful in strategic planning to determine the college's history and direction. This resource should be able to help monitor the trends in the eight critical areas of the College's strategic research emphasis.

CONCLUSION

Using computers to aid human collaboration is not a new idea. Bush's *Memex* was a vision that, although expressed without reference to computing machines, must have been inspired by the computational power he saw in Eniac [8]. Bush recognized that knowledge accessibility and the interlinking of related knowledge are central to the augmentation of human intelligence.

Metaphorically speaking, a convocation of a college faculty can be compared with a working computer: the assembly hall is a memory board, with each seat an address into which is mapped an element of associative memory known as a faculty member. The faculty in attendance are the portion of the corporate memory of the college that is "swapped in" to main memory. The numerous connections between these associative elements include friendships, acquaintances, and working relationships. In addition to these are countless *potential* relationships that might flourish, yielding novel and fruitful scientific and clinical results, but which remain undiscovered.

The use of networked information resources holds great promise for augmenting the powers of all–too–fallible human memories; the faculty research interests resource is one small way of improving the connections in the human network that is the biomedical research community. Extending the resource to include all the institutions that have contributed literature indexed in MEDLINE could create a global biomedical research network.

References

1. Humphreys B., Lindberg D., Hole W. Assessing and enhancing the value of the UMLS knowledge sources. In *15th Proc-Annu-Symp-Comput-Appl-Med-Care*, pages 78–82, Nov. 1991.
2. Fowler J., Kouramajian V., Maram S., Devadhar V. Automated MeSH indexing of the world–wide web. In *19th Proc-Annu-Symp-Comput-Appl-Med-Care*, pages 893–897, Oct. 1995.
3. Kouramajian V., Fowler J., Devadhar V., Maram S. Categorization by reference: A novel approach to automated mesh indexing. In *19th Proc-Annu-Symp-Comput-Appl-Med-Care*, pages 878–882, Oct. 1995.
4. Nelson S. J., Tuttle M. S., Cole W. G., et al. From meaning to term: Semantic locality in the UMLS metathesaurus. In *15th Proc-Annu-Symp-Comput-Appl-Med-Care*, pages 209–213, Nov. 1991.
5. Chute C. G., Yang Y., Evans D. A. Latent semantic indexing of medical diagnoses using UMLS semantic structures. In *15th Proc-Annu-Symp-Comput-Appl-Med-Care*, pages 185–189, Nov. 1991.
6. Yang Y., Chute C. G. An application of expert network to clinical classification and MEDLINE indexing. In *18th Proc-Annu-Symp-Comput-Appl-Med-Care*, pages 157–161, Nov. 1994.
7. Schwartz M. Internet Resource Discovery at the University of Colorado. In *IEEE Computer*, volume 26, pages 25–35, Sept. 1993.
8. Bush V. As we may think. *The Atlantic Monthly*, 176(1):101–108, July 1945.